

# Deriving a SNOMED CT Data Model

Ming Zhang<sup>2</sup>, Jon Patrick<sup>1</sup>, Donna Truran<sup>2</sup>, Kerry Innes<sup>2</sup>

<sup>1</sup>School of Information Technologies

University of Sydney, NSW, Australia

<sup>2</sup>National Centre for Classification in Health

University of Sydney, NSW, Australia

## ABSTRACT

*This paper identifies the underlying data model of SNOMED CT (SCT) derived through reverse engineering its distribution data tables, rather than relying on the published descriptions which may be at variance with the actual data. We need an accurate, content verified, description of the conceptual data model to compare to its hypothetical design model as a quality check. The documentation such as the "modellers style guide" that is intended to explain the workings of SNOMED is not generally available. Managing and implementing the SCT data in a conventional persistent store by third parties also requires this knowledge to maintain consistent interoperability. The analysis of the resultant data model shows that the SCT data released to third parties does not have integrity constraints, apart from identifier uniqueness, and thereby any information systems implementation will have to create its own integrity maintenance.*

## INTRODUCTION

The importance of terminological systems, classifications and ontologies, has significantly increased in recent years. Many different systems have been developed and implemented such as ICD-10, GALEN, UMLS and SNOMED CT (SCT). The International Classification of Diseases (ICD), maintained by the World Health Organization, is perhaps the best known classification system used in medicine. The Unified Medical Language System (UMLS), implemented by the National Library of Medicine, combines a large number of distinct terminologies into a single platform [1] [2]. The GALEN Common Reference Model (CRM) comprises elementary clinical concepts, relationships and complex concepts [3]. One study has been made to evaluate the consistency of SNOMED CT with the principles of ontologies such as subsumption and hierarchical structure according to a Description Logic perspective [4]. Also the forms of terminology systems have been classified from an ontology perspective and a future direction towards producing more coherent terminologies have been recommended [5].

Substantial effort has been made recently for a precise understanding of existing terminology systems and for applying them into operational information system applications. Such an application is necessary to assess whether a terminology system is appropriate for use in certain circumstances, or

when one has to design or develop a new system [6]. Various terminology systems were compared from the aspect of typology and evaluated from the perspective of conceptual model and formal representation of structure [6] [7]. On the other hand, there are many real applications which discuss issues of procedures for the integration of terminology systems into information systems. A methodology for developing and implementing, and the principles for maintaining the structure of a large expanding terminology system was researched to provide a way to keep consistency between those systems and their applications [8]. The Veterans Health Administration in United States has evaluated the coverage of the SCT terms and applied it to their information system [9].

## SNOMED CT DATA SYSTEM

SNOMED Clinical Terms provides "a common language" for use by clinical specialists in clinical notes to serve data capture purposes, retrieval and aggregation of health data. The basic features of SCT can be described as a concept-based terminology, where a clinical idea is uniquely identified and described, often with multiple descriptions. All descriptions for the same concepts are linked together. All concepts are organized into a hierarchy and, the relationships between any two concepts give information that describes characteristics of concepts.

SCT can be used in a range of information systems and is often reorganised for local needs of processing and functionality. The purpose of this paper is to describe the reverse engineering of the SCT data files, as released to third parties, to create a conceptual data model of its organisation in an EER model with the long term view of developing methods for managing and implementing the SCT data in a conventional persistent store. This process is also important as there is limited documentation available for describing the data model, there are still many debates about the design and consistency of SCT, and the data is in constant flux with new releases every six months. Such issues are particularly relevant to researchers who are only just coming to work on SCT. This study also provides some potential recommendations for updating SCT when in an existing terminology store to make it better suit local requirements. The ultimate target is to produce a data model and a conceptual model that are congruent using good data modelling principles.

An inferred data model as such will enable us to understand the extent of deviations from an underlying ideal superstructure and make decisions as to whether they form true logical extensions or are just idiosyncratic, for example some logical patterns have 1:1 mappings for 99% of their instances and the other 1% having 1:Many mappings. Understanding the intrinsic constraints in the data model are a key element of successful implementation but they are entirely implicit and hence not readily recognisable or comprehensible from the distribution data tables. The only method of unravelling the content of this type is to reverse engineer the data model from the data instances. Seasoned SNOMED researchers may find it unnecessary to perform such a task, but it is essential knowledge for any individual or group who has to build a terminology server de novo. Further, this will also be essential knowledge for users who wish to use it in an operational hospital information system, in any way more extensive than just using SCT as a lexicon of convenience, as we see in most existing products available in the market place. The point of our motivation is that further specification, constraint and documentation of the full design, power, utility and performance of SCT is required BEFORE any system designers and implementers will be able to do SNOMED CT justice. It is our hope that this study will provide an easier path for other researchers who wish to take up an interest in using SCT for significant knowledge engineering tasks, although we do not address that question directly in this paper.

## METHODOLOGY

The research work presented by this paper primarily focuses on constructing a conceptual data model for SCT from its own data. The method to fulfil this objective consists of three steps. All analyses are based on SCT version release date January 2006.

### *Data preparation*

SCT is distributed as text files transformed by the distributors in some unspecified manner from their original Protege source. Hence this distribution data needs to be restored to a framework to identify SCT's elementary data structures. Also a fast and consistent way to query and manipulate the data through the computational framework is needed. Hence, the raw data, consisting of the three tables; concepts, descriptions and relationships, were placed into the relational database management system, MySQL.

### *Investigation Structure*

To create the description of a conceptual data model represented in the SCT data, both the explicit characteristics and implicit characteristics of the data were identified. The explicit characteristics were detected by direct querying of the prepared relational database with various filters. Then, inferring from the explicit results, deductions were made to decide

which implicit structural aspects inherent and hidden in the tables might be discovered by programming more complex queries.

### *Data modelling*

After structural features were discovered from the raw data tables, conceptual data modelling of entities and the relationships was performed manually using the Extended Entity Relationship (EER) model.

## EXPERIMENTS AND RESULTS

### *Explicit characteristics - Concepts, Descriptions and relationships*

Each concept in SCT has a unique identifier and a fully specified name describing the nature of the concept. Any concept can have more than one description (ranging from 2 to 44).. SCT has 368,590 concepts and 1,014,183 descriptions. Each concept has a status which indicates whether it is in active use or the reason for its inactive status. This study only uses active concepts.

The Relationships table specifies the relationships between concepts, and the type of each relationship. Each record in the relationships table consists of a source concept and a target concept as well as the relationship type. Data retrievals show there are 65 different relationship types. The relationships table has a column for a characteristic type attribute which indicates one of four roles for the relationship record of which only two are relevant here. The first is that the target concept is used to define a source concept. For example, "burn of elbow" has the relationship "finding site" with "elbow structure". The second role is that the target concept gives optional information for qualifying the source concept. For example, the concept "appendicitis" can have the relationship "onset" with "sudden onset" and "gradual onset". So the relationship type "onset" is applied to qualify the concept. Some relationship types have more than one characteristic type according to their role in relating the target and source concepts.

The "IS\_A" relationship type organizes all concepts into a hierarchical structure of sub- and super-type concepts. An analysis of this relationship type shows all concepts traced through the "IS\_A" relationship have the same root, "SNOMED CT CONCEPT".

### *Implicit characteristics - Classification Principles*

A classification is an arrangement of objects or concepts into groups of concepts, called classes organised in a hierarchical manner based on their essential characteristics. However, SCT is not clearly a classification due to multiple inheritances in its IS\_A hierarchical structure. For instance, one concept which has more than two parents positions this concept into several classes which is not permitted in a normal classification scheme. The root concept "SNOMED CT CONCEPT" has 19 direct children called "top categories". If SCT uses a

Top category	Frequency	Top category	Frequency
Physical force	173	Context-dependent categories	4538
Special concept	446	Observable entity	7626
Specimen	1047	Qualifier value	8339
Staging and scales	1108	Pharmaceutical / biologic product	19632
Linkage concept	1131	Substance	23320
Events	8432	Organism	26436
Environments & geographical locations	1705	Body structure	31811
Physical object	4363	Procedure	53028
Social context	5195	Clinical finding	107625
Record artifact	174	<b>TOTAL (Active concepts)</b>	306129

Table 1. Active concept frequency distribution across the top 19 categories.

Source concept	Relationship type	Target concept	Number of instances
Clinical finding	Causative agent	Pharmaceutical / biologic product	510
Clinical finding	Course	Qualifier value	67792
Clinical finding	Due to	Clinical finding	1569
Clinical finding	Episodicity	Qualifier value	67528
Clinical finding	Finding site	Body structure	86140
....			

Table 2. Some valid SCT concept-relationship patterns for “clinical finding”.

classification principle, all concepts should belong to one category only. The routes for each concept in the hierarchical structure to its top category were identified. The results show (Table 1) that each concept belongs to one category only, and more than 30% of concepts are classified into the “clinical finding” top category.

#### *Inferred relationship patterns*

The above results reveal the interaction of explicit and implicit characteristics of SCT and that more implicit features can be discovered. The relationship table records any relationship link between two concepts. The format of each record is that one source concept has a relationship with a target concept. For example, the concept “breast cancer” has the relationship “finding site” with another concept “breast structure”. In classification principles, “breast cancer” belongs to the top category of clinical finding and the “breast structure” is a concept in the body structure superclass. As all concepts are classified into 19 top categories and there are 62 relationship types between all concepts, we suppose that there is restricted use of specific relationship types between any two top categories. The foregoing example represents that the “clinical finding” has a relationship of “finding site” with “body structure”. This can be treated as a record with a pattern form “top category – relationship type – top category”. With 19 top categories and 65 relationship types, the number of different patterns is  $19 \times 19 \times 65 = 23,465$ . Analysis of the data tables shows that of the 23,465 patterns, only 110 patterns have instances in the relationship table. Furthermore 19 patterns are

“IS\_A” relationships. Hence only 91 patterns are useful for analysis of the constraints of the relationships between the top categories.

## **DATA MODELLING**

The inferred 91 relationship patterns between top categories and relationship types can be modelled to generate a conceptual model for SCT’s data contents. To construct an EER data model, the top 19 categories are defined to be entities and the 91 valid patterns are the relationships between those entities. Therefore, for example, the relationships between the entity class “clinical finding” and other entity classes can be determined from the valid patterns in the Table 2. From this information a part of the model has been generated and is shown in figure 1.

The work proceeded to create an EER diagram for all entities and relationship results in a diagram that needs to be reproduced at an A3 scale to be readable (appendix 1). The data model reveals a number of telling truths that create issues for further implementations by third parties. Firstly, if such a party were to use the general rules for logical database design there would be 19 tables for the entities and 91 tables for the concept-relationship patterns, a total of 110 in all. This is to be compared to SCT’s own model of 3 tables. Secondly, it is apparent that there are no integrity constraints between any relationships in the release data, in the database sense of the term, that is, there are no optionality/mandatory constraints and all cardinality is many-to-many, creating a completely unconstrained relational model. There is the single

constraint of a unique identifier for each entity. Hence in principle database management technology could offer none of its usual features for managing the integrity of the data making for difficulties in managing the data for re-use in a conventional information system.

There is no doubt that SCT has integrity constraints within its storage format in Protégé, however those constraints are not available to third parties using the released data tables. This is a limiting condition as in order for SCT to be implemented in any sort of coherent standardized fashion the distribution format should allow the rules, and the model itself to be either discoverable through analysis of this sort – or fully documented. Otherwise every implementation in every different system will do it differently. In fact it is likely that vendors doing a full implementation of SCT would have to customise it for specific use and purposes, and without fully understanding all the content, its rules, constraints etc, such customization will result in disparate implementations between vendors which could then challenge interoperability.

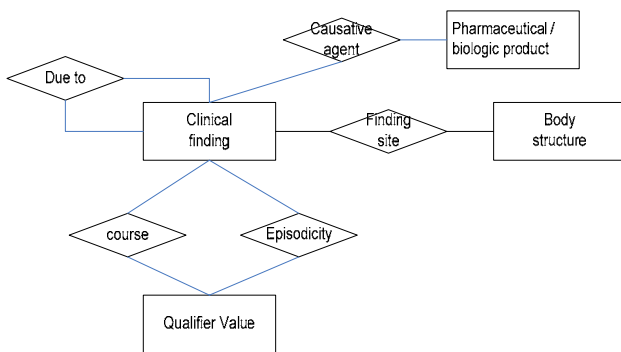


Figure 1. Part of the EER model for clinical finding.

## CONCLUSIONS

An EER conceptual data model of SCT has been inferred from both explicit and implicit characteristics of the released data. It uses the main concept classes in SCT and the internal connections between those classes. The advantage of constructing such a model is that it delivers an overview as well as providing a direct view into the interior of SCT's data configuration and gives a quality check of data against model. From the perspective of database design, the model is a useful representation to develop an alternative approach for delivering the terminology system in a robust repository rather than in the current loose spreadsheet-like format. At the moment this investigative reverse engineering is incomplete without investigating structures deeper in the hierarchy and without effective representation of multiple parents and multiple relationships between concepts. But an extension to this study, which involves a reverse engineered analysis of the SCT refined model using semantic tags (or suffixes), has already begun, and early results from this continuing study reveal that there are some cardinality constraints within SCT data which can be discovered,

and which may prove useful knowledge for implementers. Further investigation of SCT approaches to role-group relationships and post co-ordination is also necessary and planned

The analysis presented here, though performed using the EER modelling paradigm on the SCT top categories, reveals only one integrity constraint usable within a relational database model, entity integrity. Therefore, any information systems implementation will have to create its own integrity maintenance. This creates the likelihood that differences between implementers can be expected, and modifications to the SCT releases cannot be assured of consistent implementation. These early findings also indicate that considerable investigation is required to delineate the boundaries between terminology models and systems, and information models, and this is essential pre-requisite knowledge for building and maintaining interoperability.

## REFERENCES

- [1] Humphreys BL, Lindberg DA, Schoolman HM, Barnett. The Unified Medical Language System: An informatics research collaboration. *J Am Med Inform Assoc.* 1998 Jan-Feb; 5(1): 1-11.
- [2] UMLS <http://www.nlm.nih.gov/research/umls/> (2006)
- [3] OpenGALEN. <http://www.opengalen.org>. (2006)
- [4] Bodenreider O, Smith B, Kumar A, Burgun A.: Investigating subsumption in DL-based terminologies: A Case Study in SNOMED CT. *KR-MED 2004:* 12-20
- [5] Smith B, Ceusters W. "Ontology as the Core Discipline of Biomedical Informatics: Legacies of the Past and Recommendations for the Future Direction of Research", in: Dodig Crnkovic Gordana, Stuart Susan (eds.): *Computing, Philosophy, and Cognitive Science*, Cambridge Scholars Press, Cambridge, forthcoming
- [6] de Keizer NF, Abu-Hanna A, Zwetsloot-Schonk JH. Understanding terminological systems. I: Terminology and typology. *Methods Inf Med* 2000 Mar; 39(1):16-21.
- [7] de Keizer NF, Abu-Hanna A. Understanding terminological systems. II: Experience with conceptual and formal representation of structure. *Methods Inf Med.* 2000 Mar; 39(1):22-9.
- [8] Fielding JM, Simon J, Ceusters W. and Smith B. 2004 "Ontological Theory for Ontological Engineering: Biomedical Systems Information Integration", *Proceedings of the Ninth International Conference on the Principles of Knowledge Representation and Reasoning (KR2004)*, Whistler, BC, 2-5 June 2004, 114-120.
- [9] Penz JF, Brown SH, Carter JS, Elkin PL, Nguyen VN, Sims SA, Lincoln MJ. Evaluation of SNOMED-CT Coverage of Veterans Health Administration Terms. *Medinfo.* 2004:540-544.

